Emotions classification on face images

STANFORD UNIVERSITY

Mikael Jorda, Nina Miolane

mjorda@stanford.edu, nmiolane@stanford.edu

Abstract

In Robotics, emotions classification can be used to enhance human-robot interactions since the robot is capable of interpreting a human reaction. It can also be useful for research on behaviors on social networks. Here, we present a classifier of emotions on face images, trained on a small database. This classifier has the particularity to be trained on a low number of images. Given this problematic, we compared the accuracy of the traditional methods (Fisher vectors, Bag of words) used to work on small data, with the new deep methods usually working on big data samples. Our results show that with only a few images, we can achieve an accuracy of around 85% for the classification on 7 classes of emotions using different types of preprocessing on the images, and a SVM multi-class algorithm with Spatial Pytamid Kernel. We also show that the use of neural networks do not allow us to best this score on such a small database.

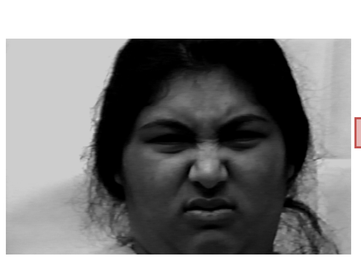
Data set

We use the Expanded Cohn-Kanabe (CK+) data base, which comprises 327 images 640 x 400 pixels, of 123 different subjects, each labelled with one of the following seven emotions: anger, contempt, disgust, fear, happiness, sadness and surprise.

Surprise

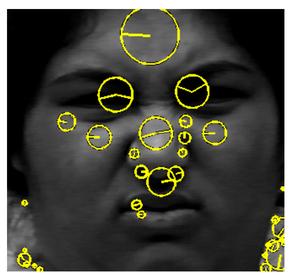
Happiness

Traditional methods



Raw image

Pre-processing



Local descriptors

Fisher vector → Train SVM Classifier With Spatial Pyramid Kernel

Bag of words → Train SVM Classifier With Spatial Pyramid Kernel

1. Preprocessing and extraction of local descriptors: In a first step, the images are cropped and intensity-normalized. A set of local descriptors are extracted, using two different methods: SIFT and dense SIFT. [3]

- Scale-Invariant Feature Transform (SIFT) is an algorithm in computer vision to detect and describe features in images. It is particularly useful to describe an image because the local descriptors extracted with this algorithm are :
  - Local and based on the appearance of the object at particular interest points
  - Invariant to affine transformations
  - Robust to changes in illumination, noise, and small changes in view point
  - Highly distinctive
  - Easy to extract

Therefore, they can render differences in the facial expression in two images, even if these are two pictures of the same subject, and they will be able to capture the facial expression in a picture even if it is darker/lighter or if it was taken with some angle.

- Dense SIFT (dSIFT) consists in placing a grid on the image and computing a SIFT descriptor at each point of the grid, instead of letting the algorithm decide where to place descriptors and compute them.

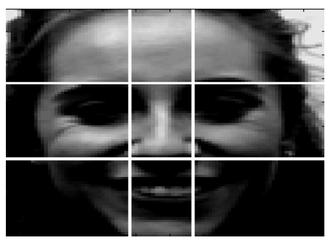
2. Featurization: We have tested two approaches for the featurization, each working either with SIFT or dense SIFT local descriptors.

- Bag-of-Words representation:
  - Learn a K-means of SIFTs (visual vocabulary)
  - Assign to each image the histogram of assignments of its SIFTs to the visual words
  - Normalizations [2]
- Fisher Vector representation:
  - Learn a GMM model of the SIFTs [3]
  - Assign to each image the Fisher Vector associated to this GMM [1,2]
  - Normalization [2]

These methods are usually used for object recognition (or face recognition). However, they can easily be generalized to emotion classification, the principal problem being to recognize the emotion instead of the resemblance between two images. On our small dataset, errors tend to happen more often when we have images of the same subject in the training and test set (with different emotions). However, when we test an image of someone we had never seen, the algorithm classifies the emotion quite precisely. This problem would be solved if we had more subjects in the database.

3. The Spatial Pyramid Kernel

- Divide each image in 9 regions
- Compute the concatenation of BoW or FV representation for each region (Mercer Kernel)



Spatial Pyramid Kernel: The final feature vector is the concatenation of the feature vector in each of the 9 zones. It amounts to running SVM with a Mercer Kernel called the Spatial Pyramid Kernel.

4. Classifier and Results

We used "one vs the rest" strategy for multi-class classification. More precisely, we train 7 classifiers to separate one emotion from the rest, solving the primal problem for L2-loss, with L2-regularization. The results below show the 10-fold cross validation accuracy on our whole dataset

Bag of words (K=400)	SVM: 75.94%
	SVM with SP Kernel: 84.71%
Fisher vector (M=30)	SVM: 78.59%
	SVM with SP Kernel: 83.79%

Average Accuracy: Comparison of the two featurization techniques. Significant increase in performance with the use of the Spatial Pyramid Kernel.

Deep convolutional neural networks

Overview

Lately, deep convolutional neural networks (CNN) have been shown to be extremely powerful for image recognition, namely for faces recognition [4]. However, these networks require a huge amount of training data in order to be trained properly and to be efficient. Therefore, our idea was to use a pre trained neural network. More precisely, we take the CNN form [4], pre trained on 2.6M images, and we cut the last layer. Then, we process our images with this cut neural network and use the resulting vector (normalized) as our feature vector. Finally, We run linear SVM with this feature vector.

Results

On our small dataset, this technique gives a 10 fold cross validation accuracy of 62% for emotion recognition. This method is less efficient than the "conventional ones". This is probably due to the face that our dataset is very small, and the neural network was trained for face recognition, and not emotion recognition.

References: [1] Sanchez, Perronnin, Mensink, Verbeek, Image Classification with the Fisher Vector: Theory and practice (2013), [2] Perronnin, Sanchez, Mensink, Improving the Fisher Kernel for Large Scale Image Classification, (2010)